J Sign Process Syst
DOI 10.1007/s11265-008-0215-5

Balancing the Role of Priors in Multi-Observer Segmentation Evaluation

Yaoyao Zhu · Xiaolei Huang · Wei Wang · Daniel Lopresti · Rodney Long · Sameer Antani · Zhiyun Xue · George Thoma

Received: 21 January 2008 / Revised: 31 March 2008 / Accepted: 12 April 2008

Abstract Comparison of a group of multiple observer segmentations is known to be a challenging problem. A good segmentation evaluation method would allow different segmentations not only to be compared, but to be combined to generate a "true" segmentation with higher consensus. Numerous multi-observer segmentation evaluation approaches have been proposed in the literature, and STAPLE in particular probabilistically estimates the true segmentation by optimal combination of observed segmentations and a prior model of the truth. An Expectation–Maximization (EM) algorithm, STAPLE's convergence to the desired local minima depends on good initializations for the truth prior and the observer-performance prior. However, accurate modeling of the initial truth prior is nontrivial. Moreover, among the two priors, the truth prior always dominates so that in certain scenarios when meaningful observer-performance priors are available, STAPLE can not take advantage of that information. In this paper, we propose a Bayesian decision formulation of the problem that permits the two types of prior knowledge to be integrated in a complementary manner in four cases with differing application purposes: (1) with known truth prior; (2) with observer prior; (3) with neither truth prior nor observer prior; and (4) with both truth prior and observer prior. The third and fourth cases are not discussed (or effectively ignored) by STAPLE, and in our research we propose a new method to combine multiple-observer segmentations based on the maximum a posterior (MAP) principle, which respects the observer prior regardless of the availability of the truth prior. Based on the four scenarios, we have developed a web-based software application that implements the flexible segmentation evaluation framework for digitized uterine cervix images. Experiment results show that our framework has flexibility in effectively integrating different priors for multi-observer segmentation evaluation and it also generates results comparing favorably to those by the STAPLE algorithm and the Majority Vote Rule.

Keywords Ground truth · Bayesian decision · Precision · Segmentation · Multi-observer · Sensitivity · Specificity · STAPLE · Validation

Y. Zhu (✉) · X. Huang · W. Wang · D. Lopresti
Department of Computer Science and Engineering,
Lehigh University,
Bethlehem, PA 18015, USA
e-mail: yaz304@lehigh.edu

X. Huang
e-mail: xih206@lehigh.edu

W. Wang
e-mail: wew305@lehigh.edu

D. Lopresti
e-mail: dal9@lehigh.edu

R. Long · S. Antani · Z. Xue · G. Thoma
National Library of Medicine, National Institutes of Health,
Bethesda, MD 20894, USA

R. Long
e-mail: rlong@mail.nih.gov

S. Antani
e-mail: santani@mail.nih.gov

Z. Xue
e-mail: xuez@mail.nih.gov

G. Thoma
e-mail: gthoma@mail.nih.gov

1 Introduction

Segmentation is a fundamental problem in many pattern recognition and image processing applications. Segmenta-

tions can be generated by different automated computer methods or by human observers. Multiple-observer segmentation evaluation is helpful in many scenarios. Some examples are: (a) evaluating performance of multiple observers' segmentations simultaneously [1]; (b) measuring segmentation complexity [2]; (c) combining multiple observers' segmentations to generate the ground-truth segmentation. STAPLE [1] is an algorithm proposed for the first scenario.

In STAPLE, two different kinds of prior knowledge can be integrated. One is the *truth prior*, which specifies the probability of each pixel being inside the segmentation. This information can be obtained through training a statistical atlas. The other is the *observer-performance (or observer) prior*, which specifies prior knowledge about the performance level of each observer, often quantified by two performance parameters, *sensitivity* and *specificity* (Section 2.2). However, the role of the two priors is not balanced in the STAPLE algorithm. The truth prior is heavily depended on, and it almost always dominants over the observer prior so that the observer prior has little effect on the final evaluation result. Since the truth prior is often unknown and an estimated prior is used instead, the evaluation result is often not in agreement with the initial performance measures of observers. As pointed out in [1], if this was discovered in the application, it would indicate either the need to re-evaluate the global prior assumption or the need for improved training of the experts generating the segmentations. However these recommendations do not address the lack of truth prior or the discrepancy caused by inconsistent truth and observer priors. In certain situations, the performance measures of multiple observers' segmentations are known in advance to some extent. For instance, let us consider an observer which is an automated segmentation algorithm, we will know if the algorithm tends to perform conservatively thus has a low specificity. For manual segmentations, we can assume that segmentations made by experts have higher sensitivity and specificity than those by non-experts. In these situations, we would desire evaluation results that are consistent with the known observer-performance priors.

Based on the above observations, we propose a different framework based on the Bayesian Decision Theory and the MAP optimization principle for the multiple-observer segmentation evaluation problem. The framework is based on different segmentation evaluation needs and different prior knowledge available. One need is to estimate the ground-truth segmentation and observer performance levels, with or without the truth prior probability. The other need is to combine the segmentations from observers with different measures of performance. To address the first need, if the truth prior is unknown, the observers are treated equally as experts with high sensitivity and specificity. The truth prior probability is estimated by averaging all observer segmentations then integrated in the MAP estimation. If a reliable truth prior is available, it will be used directly. To address the second need where we know *a priori* some observers' sensitivity and specificity, the MAP solution combines these performance measures to compute a ground truth map which is consistent with the known measures. The estimated ground truth can then be used to evaluate other observers whose performance measures are unknown. For validation purposes, gold-standard ground truth segmentation can be acquired in phantom experiments or by multiple-observer consensus to compare with the estimated ground truth.

We developed an online software system to evaluate multi-observer segmentations for medical images such as those in the NCI/NLM medical repository of digital cervicographic images (cervigrams) [3]. The total 939 images were collected as part of a study for the evolution of lesions related to cervical cancer conducted by the National Cancer Institute (NCI) together with the National Library of Medicine (NLM) through two major studies in Costa Rica and the United States, the Guanacaste and ALTS1 projects, respectively [19]. In these studies, multiple observers (or raters) have marked several important regions on cervigrams that are of anatomical or clinical interest, including the cervix boundary and acetowhite regions. They were clinicians with expertise in colposcopy that were identified by members of the Board of Directors of the American Society for Colposcopy and Cervical Pathology and by staff at the National Cancer Institute. They included 12 general gynecologists and 8 gynecologist oncologists. 18 of them work in academic settings and 2 in private practice. They have varies of years of experience. In the studies, the total number of subjects was also 939 (one cervigram per subject). The cervix boundary defines the region of the uterine cervix, which is of anatomic interest within the cervigrams. The acetowhite regions are epithelium with whitened appearance, which is visible for a short period of time following the application of 3% to 5% acetic acid. Some acetowhite regions correlate with uterine cervix cancer progression, and thus are of clinical significance. Examples of these marked regions are shown in Fig. 1. Each cervigram has associated with a different number of manual markings varying from one to twenty. In this paper, we consider combining multiple observers' segmentations of the cervix boundary (yellow line in Fig. 1), and our software can be used to evaluate these multi-observer segmentations in different scenarios.

The remainder of this paper is organized as follows. In Section 2, we discuss previous work and our choice for multiple-observer segmentation measures. In Section 3, we discuss previous work on combining multiple-observer segmentations. We introduce the STAPLE algorithm [1]

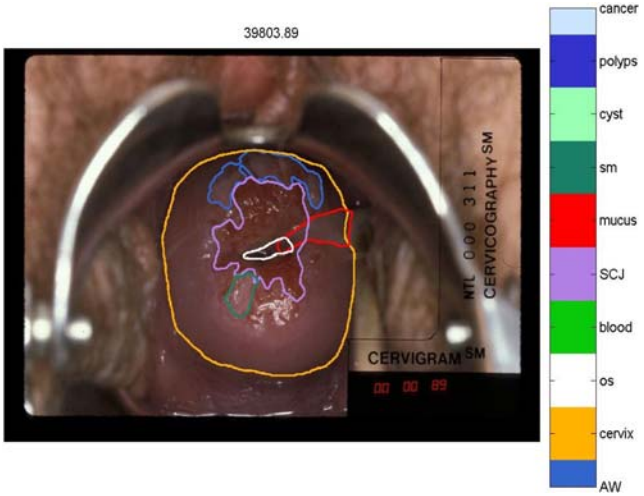


Figure 1 Several regions marked on cervigrams.

and identify its limitation. We then describe our framework and algorithms for different multiple-observer segmentation evaluation scenarios in Section 4. The web-based multiple-observer segmentation evaluation software developed based on our method is presented in Section 5, and we demonstrate experimental results and comparison with previous work in Section 6 by using the multiple observer manual segmentations. In Section 7, we also demonstrate experimental results but by using manual segmentation results to evaluate our automatic segmentation method. Section 8 concludes the paper with discussion of future work.

2 Multiple Observer Segmentation Measures

2.1 Previous Work on Segmentation Evaluation Metrics

A number of metrics have been proposed to compare segmentations. Generally the evaluation methods of image segmentation can be classified into three categories [4]: analytical methods, empirical goodness methods and empirical discrepancy methods. Analytical methods are not used to judge the performance of segmentation methods but their properties, principles, complexity, requirement and so forth. Empirical goodness methods are used to compute some manner of "goodness" criterion such as uniformity within regions, contrast between regions, shape of segmentation regions and so forth. The empirical discrepancy methods evaluate segmentation methods by comparing the segmented image against a manually segmented reference image, which is often referred as the ground truth, and computing error measures. The empirical discrepancy methods have been the most commonly used methods for segmentation evaluation.

Reviewing work in the literature, one can find two kinds of empirical discrepancy methods: (1) region-based evaluation, which evaluates segmentation consensus in terms of the number of regions, and the locations, sizes and other statistics of the segmented regions, and (2) boundary-based evaluation, which evaluates segmentation in terms of both the location and shape accuracies of the extracted region boundaries. The segmentation performance-level criteria in region-based evaluation can be: (a) sensitivity and specificity, where sensitivity is defined as "true positive fraction", and specificity is "true negative fraction" [1], (b) correctness and completeness, or precision and recall [17, 18], where high completeness means that the region segmented has covered the relevant pattern well, whereas high correctness implies that the region segmented does not contain many (incorrect) irrelevant patterns, (c) the number of misclassified pixels and their distances to the nearest correctly segmented pixels [5], (d) measures based on hamming distance between two segmentations [6], (e) local consistency error which quantifies the consistency between image segmentations of differing granularities [7], (f) bidirectional consistency error which penalizes dissimilarity between segmentations proportional to the degree of region overlap [7], and (g) partition distance which is defined as "given two partitions P and Q of S, the partition distance is

Figure 2 The relationship between sensitivity p and specificity q, and typical (p, q) values of medical experts and non-experts.

Decision \ Truth	0	1
0	q	1-p
1	1-q	p

Decision \ Truth	0	1
0	0.999	
1		0.999

Expert

Decision \ Truth	0	1
0	0.2	
1		0.2

Non-expert

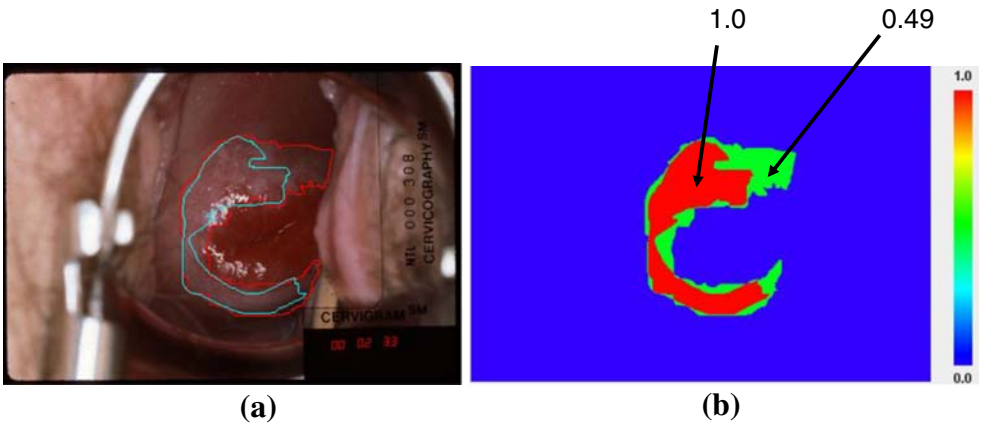


Figure 3 **a** Two observers' segmentations of acetowhite regions. **b** Multi-expert ground truth map.

the minimum number of elements that must be deleted from S, so that the two induced partitions (P and Q restricted to the remaining elements) are identical" [8]. On the other hand, the performance-level criteria in boundary-based evaluation can be: (a) distance of distribution signatures which is based on the distance between distribution signatures that represent boundary points of two segmentation masks [6], (b) precision-recall measurement which uses precision and recall values to characterize the agreement between the oriented boundary edge elements of two segmentations' region boundaries [7], and (c) a new discrepancy measure [9] which takes into account not only the consensus of the localized boundaries of the created segments but also under-segmentation and over-segmentation.

A good evaluation method would allow segmentations by different approaches not only to be compared, but to be integrated to generate segmentation with higher consensus. In our framework, we choose to use the region-based evaluation metric: sensitivity and specificity, which can be incorporated into our framework in combining multiple-observer segmentations.

2.2 Measures of Performance in Our Framework: Sensitivity and Specificity

In the NCI cervigram database, we have segmentations of regions marked by 20 observers. Since these markings can vary in size and location it is essential that we have measures to evaluate these multi-observer segmentations. We choose sensitivity p and specificity q to measure the performance level of each binary segmentation.

Sensitivity is the "true positive fraction" and defined as

$$\text{Sensitivity} = \frac{\text{TP}}{\text{TP} + \text{FN}} \tag{1}$$

where TP is the number of true positive pixels and FN is the number of false negative pixels. That is, sensitivity means the percentage of pixels properly included in the segmentation result out of all pixels in the segmentation result.

Specificity is the "true negative fraction" and defined as

$$\text{Specificity} = \frac{\text{TN}}{\text{TN} + \text{FP}} \tag{2}$$

Table 1 Configurations in the first group of experiments demonstrating that the initial (p, q) values have little effect on the result of the STAPLE algorithm.

Experiment	Global ground-truth prior γ	Value	Observer 1 (red)	Observer 2 (blue)	Estimated ground truth map
Experiment 1 (STAPLE)	Average	Initial p	0.9999	0.9999	Fig. 4b
		Initial q	0.9999	0.9999	
		Final p	0.728	0.987	
		Final q	0.972	0.862	
Experiment 2 (STAPLE)	Average	Initial p	0.9999	0.9999	Fig. 4c
		Initial q	0.9999	0.5	
		Final p	0.753	0.943	
		Final q	0.9999	0.847	
Experiment 3 (STAPLE)	Average	Initial p	0.5	0.9999	Fig. 4d
		Initial q	0.9999	0.9999	
		Final p	0.705	0.975	
		Final q	0.9999	0.876	

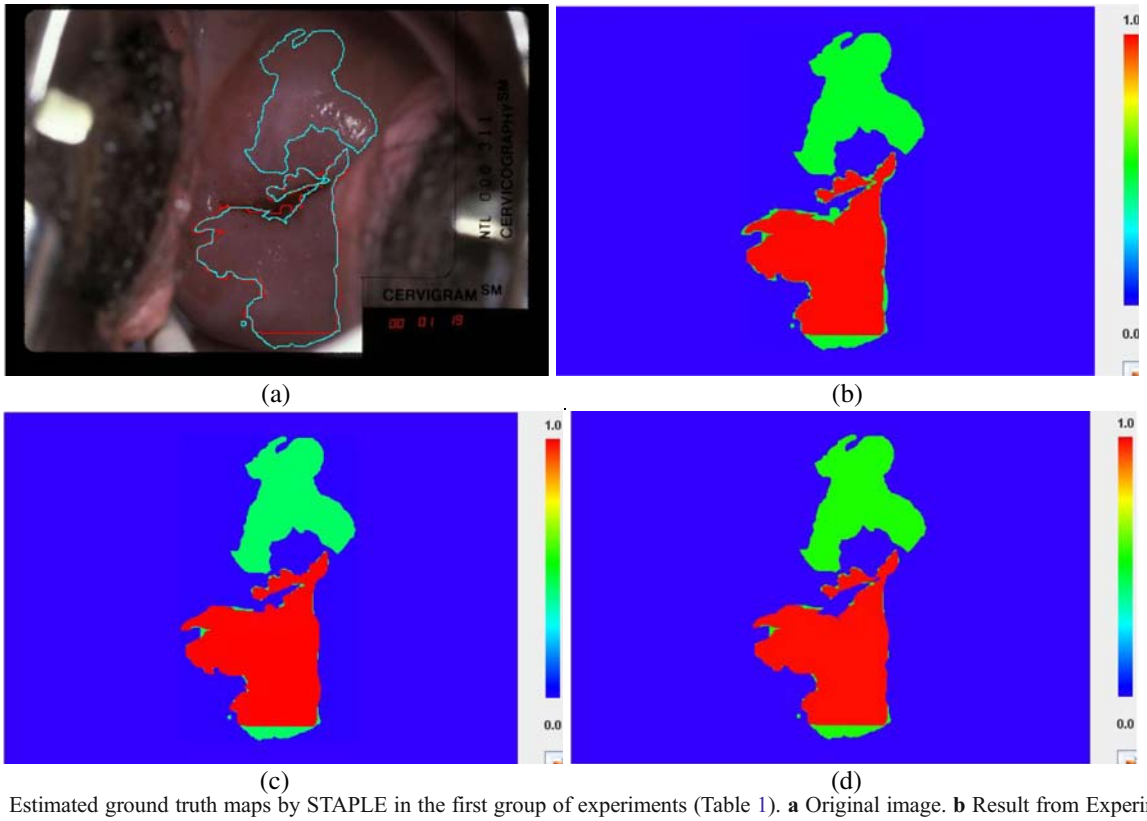


Figure 4 Estimated ground truth maps by STAPLE in the first group of experiments (Table 1). **a** Original image. **b** Result from Experiment 1. **c** Result from Experiment 2. **d** Result from Experiment 3.

where TN is the number of true negative pixels and FP is the number of false positive pixels. So specificity means the percentage of pixels properly excluded from the segmentation result out of all pixels outside of the ground truth.

The relationship of sensitivity p and specificity q in a binary segmentation can be easily understood through the diagram in Fig. 2. The pixels labeled 1 are inside the segmentation (foreground) and those labeled 0 are outside (background). Different observers may have different (p, q) values; for instance, medical experts have higher (p, q) values while inexperienced non-experts may have lower ones (Fig. 2).

The similar measures to sensitivity and specificity are correctness and completeness, or precision or recall, which are defined as follows:

$$\text{Correctness} = \frac{\text{TP}}{\text{TP} + \text{FP}}$$

$$\text{Specificity} = \frac{\text{TP}}{\text{TP} + \text{FN}}$$

Table 2 Configurations in the second group of experiments, demonstrating the dominant effect of the truth prior probability in STAPLE.

Experiment	γ	Value	Observer 1 (red)	Observer 2 (blue)	Ground truth map
Experiment 1 (STAPLE)	0.5	Initial p	0.9999	0.9999	Fig. 5b
		Initial q	0.9999	0.9999	
		Final p	0.563	0.931	
		Final q	0.9999	0.9999	
Experiment 2 (STAPLE)	0.2	Initial p	0.9999	0.9999	Fig. 5c
		Initial q	0.9999	0.9999	
		Final p	0.9999	0.9999	
		Final q	0.974	0.759	
Experiment 3 (STAPLE)	0.2	Initial p	0.5	0.9999	Fig. 5d
		Initial q	0.9999	0.9999	
		Final p	0.9999	0.9999	
		Final q	0.974	0.759	

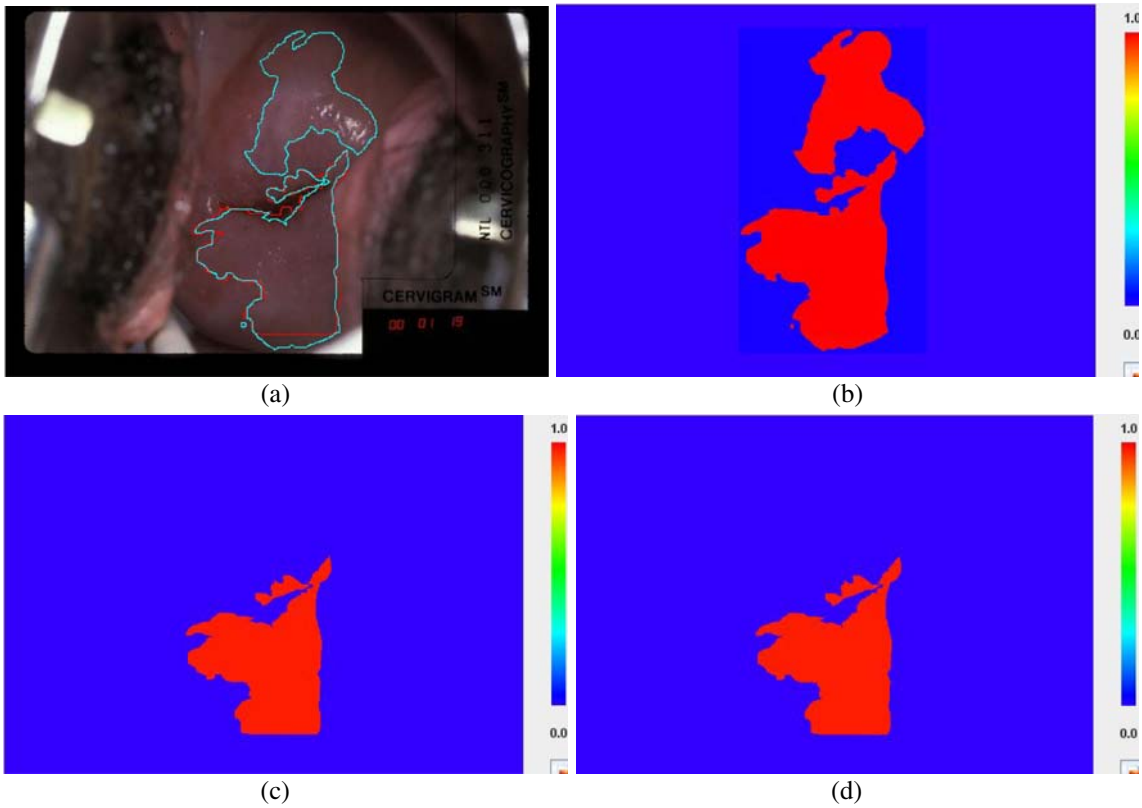


Figure 5 Estimated ground truth maps by STAPLE in the second group of experiments (Table 2). **a** original image. **b** Result from Experiment 1. **c** Result from Experiment 2. **d** Result from Experiment 3.

3 Background on Multiple Observer Segmentation Combination Methods

There are a number of combination methods proposed in the literature for different cases of integrating multi-observer segmentations to derive a final segmentation. These include class probability combining strategies such as the Min Rule, the Max Rule, the Median Rule and the Majority Vote Rule [10]. For instance, the Majority Vote Rule chooses the segmentation label for each pixel based on what the majority of observers agree on; this simple method, however, does not take into consideration the

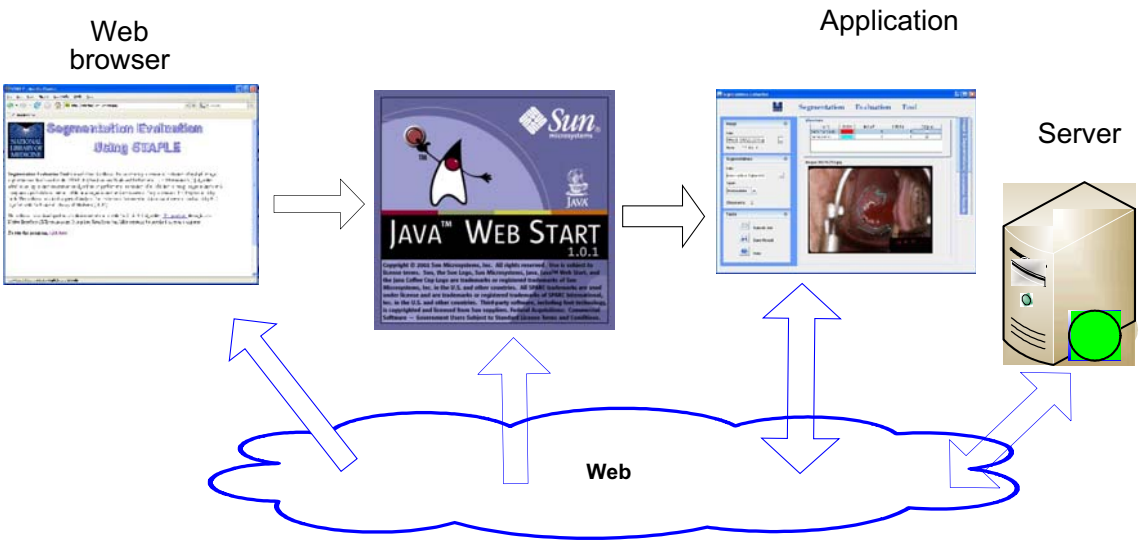


Figure 6 The architecture of the software system.

variability in quality or performance among the voters and also does not incorporate the prior knowledge regarding segmentations. There are also combination strategies that assume each classifier has expertise in a subset of the decision domain [11–13], and strategies [14, 15] that can account for different confidence or uncertainty levels in segmentations.

The result after probabilistically combining multiple observer segmentations is usually presented in a multiple-observer ground truth map. One example is shown in Fig. 3. In Fig. 3a, two observers have marked the acetowhite regions (in red line and in blue line). Figure 3b shows the corresponding ground truth map after combining the two segmentations, and in this map, each pixel is represented by a color indicating the probability that belongs inside the ground truth segmentation.

3.1 STAPLE Algorithm

The STAPLE algorithm is a well-known method proposed by Warfield et al. [1], for generating ground truth segmentation maps from the observations of multiple observers and measuring the performance levels of each of the observers.

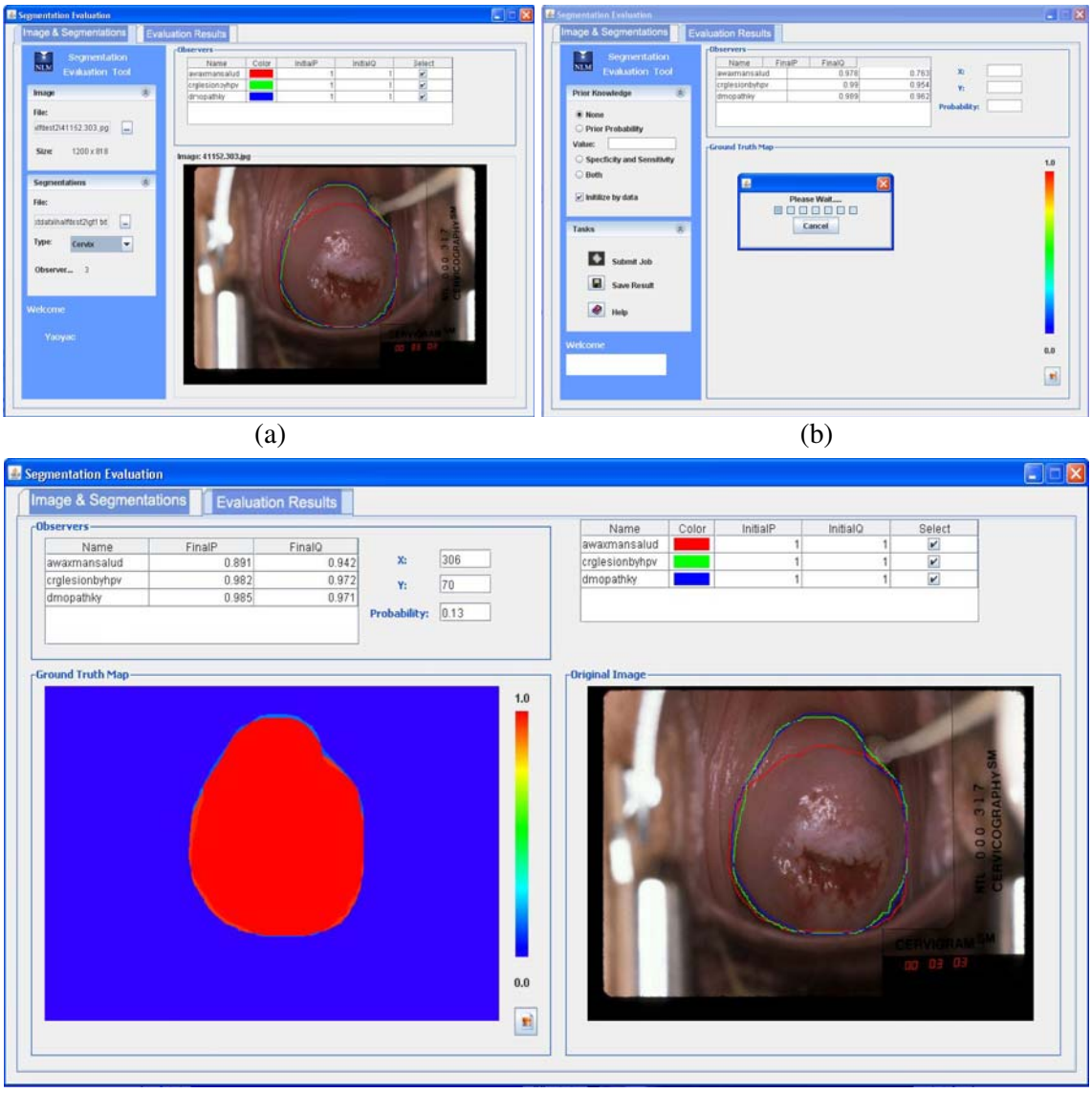
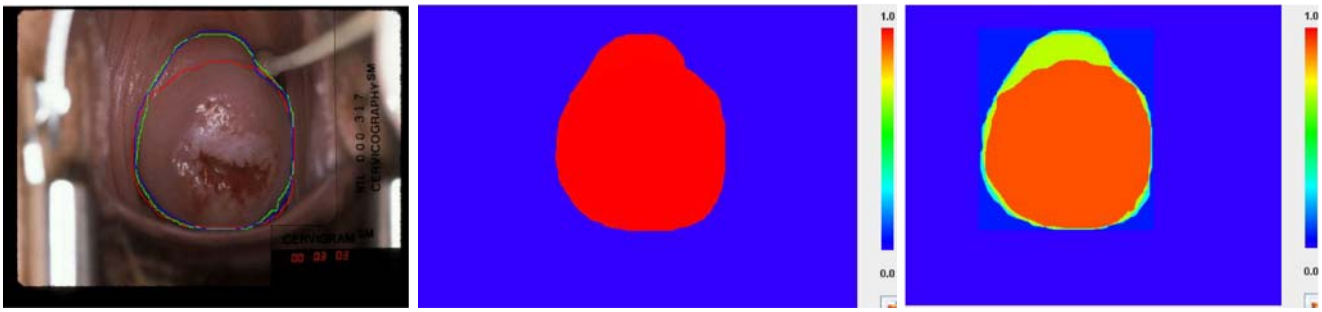


Figure 7 User interfaces of the web-based software tool. **a** Image and segmentation loading and viewing. **b** Submitting the job to the server. **c** Ground truth map and the original image.



(a) (b) (c)

Figure 8 Estimated ground truth maps with the setups in Table 3. **a** original image. **b** Result for Experiment 1. **c** Result for Experiment 2.

3.1.1 Notations

Let us suppose there are N pixels in the image whose segmentations are being evaluated by a total of R observers. The following notations are used in describing the STAPLE algorithm.

A. $\boldsymbol{p}=(p_1, p_2, ..., p_R)^T$ is a column vector of R elements, with each element a sensitivity parameter charactering one of the R segmentations;
B. $\boldsymbol{q}=(q_1, q_2, ..., q_R)^T$ is a column vector of R elements, with each element a specificity parameter charactering one of the R segmentations;
C. $\boldsymbol{D}$: an $N\times R$ matrix describing the binary decisions made for each segmentation;
D. $\boldsymbol{T}$: an indicator vector of N elements, representing the hidden binary true segmentation. For a pixel i, the structure of interest is recorded as present ($\boldsymbol{T}_i=1$) or absent ($\boldsymbol{T}_i=0$);
E. $\gamma=\mathbf{f}(\boldsymbol{T}_i=1)$, $i=1,...,$ N: the global prior probability of ($\boldsymbol{T}_i=1$), assuming equal prior probability at every pixel.

3.1.2 Algorithm

STAPLE is an EM (Expectation–Maximization) algorithm and estimates simultaneously the true segmentation $\boldsymbol{T}$ and performance level parameters of observers characterized by parameters ($\boldsymbol{p}$ and $\boldsymbol{q}$ in this case). It aims to maximize the complete data log likelihood:

$$(\widehat{p},\widehat{q}) = \arg\max_{p,q} \ln f(D,T|p,q) \tag{3}$$

Like other EM algorithms, the STAPLE algorithm has two steps: the Expectation (E) step and the Maximization (M) step. In the E step, it computes an expectation of the likelihood at each iteration k:

$$f\left(T_i|D_i,p^{(k-1)},q^{(k-1)}\right) = \frac{\prod_j f\left(D_{ij}|T_i,p_j^{(k-1)},q_j^{(k-1)}\right)f(T_i)}{\sum_{T_i'}\prod_j f\left(D_{ij}|T_i',p_j^{(k-1)},q_j^{(k-1)}\right)f(T_i')} \tag{4}$$

where the posterior probability of the true segmentation at each pixel is

$$W_i = f\left(T_i=1|D_i,p^{(k-1)},q^{(k-1)}\right) = \frac{f(T_i=1)\alpha_i}{f(T_i=1)\alpha_i+(1-f(T_i=1))\beta_i}. \tag{5}$$

In the above definition for the posterior ground truth segmentation, $f(T_i=1)$ is the truth prior probability, α_i is the conditional data probability $f(D_i=1|T_i=1,p^{(k-1)}, q^{(k-1)})$,

Table 3 Initial (p, q) values with t=0.9999 and t=0.7.

Experiment	γ	Value	Observer 1 (red)	Observer 2 (green)	Observer 3 (blue)	Result
Experiment 1	0.5	Initial p	0.9999	0.9999	0.9999	Fig. 8b
		Initial q	0.9999	0.9999	0.9999	
		Final p	0.893	0.983	0.986	
		Final q	0.946	0.971	0.969	
Experiment 2	0.5	Initial p	0.7	0.7	0.7	Fig. 8c
		Initial q	0.7	0.7	0.7	
		Final p	0.893	0.983	0.986	
		Final q	0.946	0.971	0.969	

Table 4 Truth prior probability and (p, q) values initialized with observer data.

Experiment	γ	Value	Observer 1 (red)	Observer 2 (green)	Observer 3 (blue)	Result
Experiment 1	0.710	Initial p	0.893	0.983	0.986	Fig. 9b
		Initial q	0.946	0.971	0.969	
		Final p	0.893	0.983	0.986	
		Final q	0.946	0.971	0.969	

and β_i is the conditional data probability $f(D_i=0|T_i=1, p^{(k-1)}, q^{(k-1)})$:

$$\alpha_i = \prod_{j:D_{ij}=1} p_i^{(k-1)} \prod_{j:D_{ij}=0} \left(1-p_j^{(k-1)}\right),$$
$$\beta_i = \prod_{j:D_{ij}=0} q_i^{(k-1)} \prod_{j:D_{ij}=1} \left(1-q_j^{(k-1)}\right) \quad (6)$$

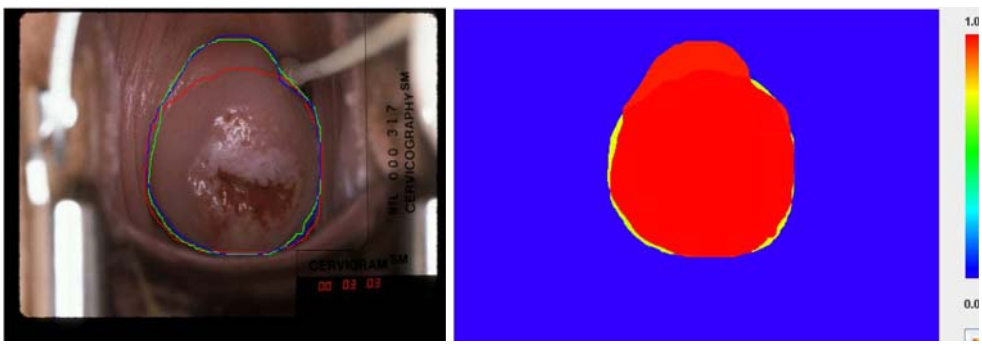
In the M step, it estimates the observers' performance level parameters, $\boldsymbol{p}$ and $\boldsymbol{q}$, that maximize the conditional expectation of the complete data log likelihood function.

$$p_j^{(k)} = \frac{\sum_{i:D_{ij}=1} W_i^{(k-1)}}{\sum_i W_i^{(k-1)}}, \quad q_j^{(k)} = \frac{\sum_{i:D_{ij}=0} \left(1-W_i^{(k-1)}\right)}{\sum_i \left(1-W_i^{(k-1)}\right)} \quad (7)$$

If the difference between (p, q) values at the k−1 and k steps is small enough, the algorithm is considered converged. Then it outputs the final (p, q) values and the ground truth map W_i. In STAPLE, there are three inputs that are needed: the multiple observer segmentations **D**, the initial (p, q) values for each segmentation, and the ground-truth segmentation prior probability $f(T_i=1)$.

3.2 Limitation of the STAPLE Algorithm

As described above, besides the multiple observer segmentation data, there are two kinds of priors that are necessary inputs to the STAPLE algorithm: the truth prior probability $f(T_i=1)$, and the observer prior represented by the initial (p, q) values for each observer's segmentation. However, as noticed by the STAPLE authors [1] and by us through extensive experiments, the truth prior almost always dominates the observer prior so that the initial (p, q) observer performance-level values have little effect on the final posterior segmentation. Indeed the converged result on (p, q) by STAPLE often contradicts the initial (p, q) prior. We believe this discrepancy is caused by the independence assumption made by STAPLE—the ground truth $\boldsymbol{T}$ is independent of the performance level parameters so that $f(\boldsymbol{T}, p, q)=f(\mathrm{T})f(p, q)$. It is obvious from the definitions of sensitivity p (Eq. 1) and specificity q (Eq. 2) that (p, q) are not independent of **T**. Having this independence assumption separates the influence of the truth prior from that of the observer performance-level prior. In practice, this manifests in a way that STAPLE can not deal with the scenario when the (p, q) values for each observer's segmentation are known. Moreover, the truth prior is often unknown and an estimated prior is used instead; if the estimated prior is far-off from the ground truth segmentation, the negative effect of the lack of prior can get magnified. This limitation of the STAPLE algorithm can be seen from the following experiments. In the first group of experiments (Table 1 and Fig. 4), the truth prior probability is not available and it is estimated as the average of the relative proportion of the labels (1 or 0) in the multiple-observer segmentations. Therefore the value of the prior probability is kept the same for all three experiments. We vary the initial (p, q) values of the two observers in different experiments: in Experiment 1, both observers are set as experts with high (p, q) values; in Experiment 2, observer 1 is set as an expert and observer 2 as a non-expert; the configurations in Experiment 3 is on the contrary to Experiment 2. Using the sample mean of the multi-observer segmentations as the truth prior, one can see



(a) (b)

Figure 9 Ground truth map with the setup of Table 4. **a** original image. **b** Result for Experiment 1.

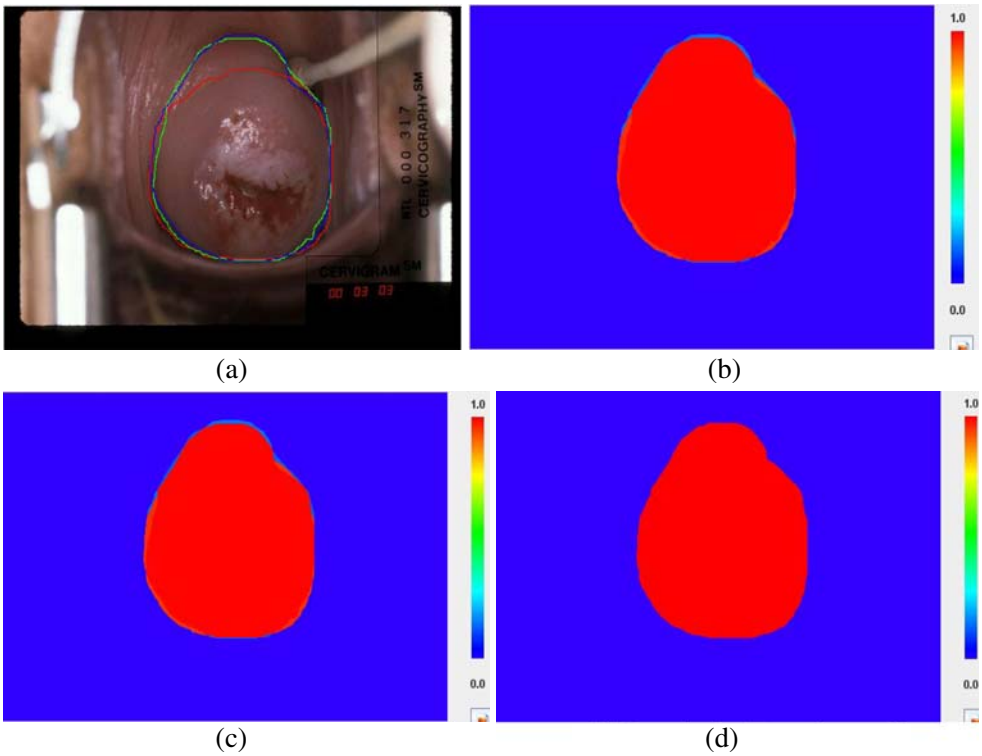


Figure 10 Estimated ground truth maps with the setups in Table 5. **a** Original Image. **b** Result for Experiment 1. **c** Result for Experiment 2. **d** Result for Experiment 3.

from Table 1 that: (1) the final estimated ground truth map by STAPLE is close to that generated by the majority vote rule, (2) the differing observer-performance prior (p, q) values have little effect on the estimated ground truth map, and (3) the converged (p, q) values can deviate greatly from the initial (p, q) values which indicates that the observer prior was overwhelmed by the truth prior (Table 1; Fig. 4).

In the second group of experiments (Table 2 and Fig. 5), we specify different truth prior probability with different (p, q) values across experiments. In Experiment 1, the truth prior probability is set to be closer to the segmentation by Observer 2 (γ=0.5), while in experiments 2 and 3, the truth prior is closer to the segmentation by Observer 1 (γ=0.2). These experiments clearly show that the truth prior probability has dominant effect on the estimated ground truth map at the converged local minima in the STAPLE algorithm. Experiment 3 in particular is interesting. In that experiment, we set Observer 1 as a non-expert and Observer 2 as an expert thus the truth prior probability is not in agreement with the prior performance measures of the two observers' segmentations. The converged results from the STAPLE algorithm are consistent with the truth prior probability instead of the observer performance-measure prior values. Experiment 3 clearly demonstrates that, even when reliable information about observer performance measures is available, we still have to get the correct truth prior in order to obtain meaningful results using STAPLE (Table 2; Fig. 5).

Table 5 The initial and final (p, q) values in the STAPLE algorithm, and majority vote rule.

Experiment	Value	Observer 1 (red)	Observer 2 (green)	Observer 3 (blue)	Result
Experiment 1 (STAPLE)	Initial p	0.9999	0.9999	0.9999	Fig. 10b
	Initial q	0.9999	0.9999	0.9999	
	Final p	0.891	0.982	0.985	
	Final q	0.942	0.972	0.971	
Experiment 2 (STAPLE)	Initial p	0.7	0.7	0.7	Fig. 10c
	Initial q	0.7	0.7	0.7	
	Final p	0.891	0.982	0.985	
	Final q	0.942	0.972	0.971	
Experiment 3 (Majority Vote Rule)	Initial p	N/A	N/A	N/A	Fig. 10d
	Initial q				
	Final p				
	Final q				

Table 6 (p, q) values for experiments of scenario two: with known p and q values for each observer.

Experiment	Value	Observer 1 (red)	Observer 2 (green)	Observer 3 (blue)	Result
Experiment 1	Initial p	0.9999	0.9999	0.9999	Fig. 11b
	Initial q	0.9999	0.9999	0.9999	
	Final p	0.893	0.983	0.986	
	Final q	0.946	0.971	0.969	
Experiment 2	Initial p	0.9999	0.9999	0.9999	Fig. 11c
	Initial q	0.9999	0.7	0.7	
	Final p	0.9999	0.981	0.984	
	Final q	0.958	0.751	0.75	

4 Problem Formalization and Algorithms in Our Framework

As demonstrated above, STAPLE effectively ignores the observer performance measure prior. Indeed in the derivation of STAPLE [1], the observer performance prior probability $f(p, q)$ was cancelled out by making the independence assumption between **T** and (p, q) values. The result of this cancellation is that there is no way to inject prior knowledge about individual observer's performance level in the STAPLE framework. Furthermore, the ground truth prior probability has shown dominant effect on the estimated posterior ground truth map and the estimated performance measures (Tables 1 and 2, Figs. 4 and 5), which is not always desirable because oftentimes we do not have reliable information about the truth prior. We argue that these limitations stem from the independence assumption because based on either the standard definitions for sensitivity and specificity (Section 2.2) or the definitions in STAPLE ($p_j = Pr(D_{ij}=1|T_i=1)$, $q_j = Pr(D_{ij}=0|T_i=0)$), p and q are fully dependent on $\boldsymbol{D}$ and $\boldsymbol{T}$. That is, given segmentation data decisions $\boldsymbol{D}$ and the ground truth $\boldsymbol{T}$, the performance measures of any Observer j, p_j and q_j, are uniquely determined.

Based on the above analysis, we propose a new framework for multiple observer segmentation evaluation, which is more general than STAPLE. We explicitly take into account different kinds of prior knowledge that are available and apply different methods in different scenarios. The two kinds of prior knowledge that can be injected into our framework are: the (ground) truth prior ($\gamma = f(T_i=1)$), and the observer performance-level prior (p, q) values. If a certain prior is unknown, it can be initialized with uniform distribution or initialized based on observers' segmentation data.

The overall theoretical framework is based on the Bayesian Decision Theory [16], which aims to make a decision based on the posterior probability distribution, $f(T|D)$. The standard *maximum a posteriori* (MAP) estimator can be applied to select the most probably ground truth **T**:

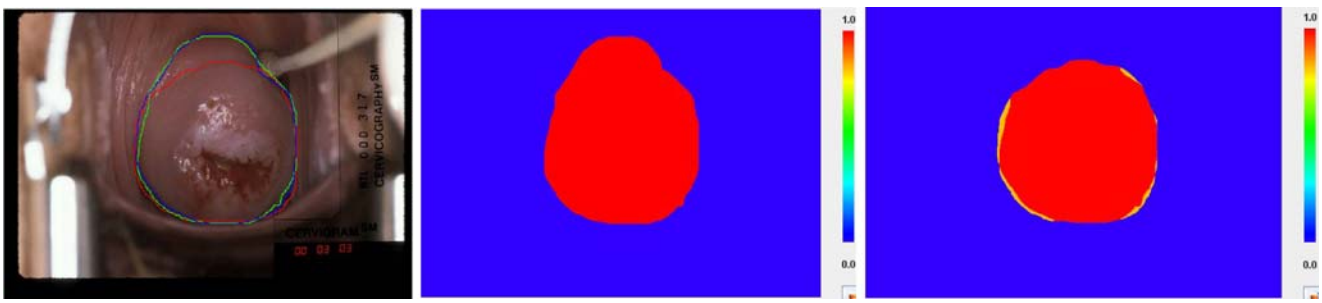
$$T^* = \arg\max_T f(T|D) \tag{8}$$

where

$$f(T|D) = \frac{f(D|T)f(T)}{f(D)} = \frac{f(D|T)f(T)}{\sum_T f(D|T)f(T)} \tag{9}$$

For pixel i, let

$$A_i = f(D_{ij}|T_i=1)f(T_i=1) = \left(\prod_{j:D_{ij}=1} p_j \prod_{j:D_{ij}=0}(1-p_j)\right) f(T_i=1) \tag{10}$$



(a) (b) (c)

Figure 11 Experimental results for scenario two: with known p and q for each observer. **a** Original Image. **b** Result for Experiment 1. **c** Result for Experiment 2.

Table 7 The (p, q) values in the experiments for the case with known p and q values for some observers.

Experiment	Value	Observer 1 (red)	Observer 2 (yellow)	Observer 3 (blue)	Observer 4 (purple)	Result
Experiment 1	p	0.9999	0.9999	0.991	0.999	Fig. 12b
	q	0.9999	0.9999	0.707	0.681	
	Known	Yes	Yes	No	No	
Experiment 2	p	0.884	0.923	0.9999	0.9999	Fig. 12c
	q	0.877	0.966	0.9999	0.9999	
	Known	No	No	Yes	Yes	

$$B_i = f(D_{ij}|T_i=0)f(T_i=0) = \left(\prod_{j:D_{ij}=0} q_j \prod_{j:D_{ij}=1}(1-q_j)\right) f(T_i=0) \tag{11}$$

Combining Eqs. 9, 10 and 11, we have:

$$f(T_i=1|D) = \frac{f(D|T_i=1)f(T_i=1)}{\sum_{T_i} f(D|T_i)f(T_i)} = \frac{\mathrm{A}_i}{\mathrm{A}_i + \mathrm{B}_i} \tag{12}$$

where $f(T_i=1|D)$ indicates the posterior probability of the true segmentation at pixel i being equal to one. It follows that the posterior background probability $f(T_i=0|D)=(1-f(T_i=1|D)$. Thus the MAP estimator (Eq. 8) will assign the class label of pixel i to be 1 (i.e. foreground pixel, $T_i=1$) if $f(T_i=1|D)>0.5$, or assign the label 0 (i.e. background pixel, $T_i=0$) if $f(T_i=1|D)<0.5$.

Next we discuss several scenarios with different prior knowledge available and different application purposes. The multi-observer segmentation evaluation algorithms in our framework are introduced for each scenario.

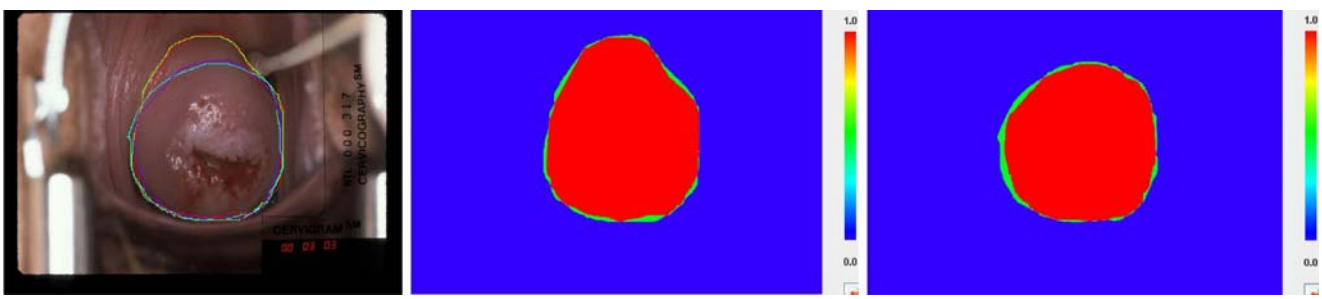
4.1 Both Truth Prior Probability $\gamma=f(\mathrm{T_i}=1)$ and Observer (p, q) Values are Known

In this scenario, we simply apply Eqs. 10, 11 and 12 with these numbers to calculate $f(T_i=1|D)$ and estimate the posterior ground truth segmentation map. This case can not be handled by STAPLE because the observer prior would be ignored and would not have the desired effect on the estimated ground truth segmentation.

4.2 Only Observer (p, q) Values are Known

In this scenario, we know the sensitivity and specificity of each observer thus we can distinguish observers of different performance levels such as experts vs. non-experts. However, we do not know the truth prior probability $f(T_i=1)$. In practice, such a situation is quite common. The sensitivity and specificity for each observer can be estimated based on training data from the observer's past experience (manual segmentations). Or if an observer is an automated segmentation algorithm, the (p, q) values of the observer can be estimated based on the characteristics of the segmentation algorithm or based on its performance on validation datasets. In this case, we want to obtain the ground truth consistent with the known (p, q) values of observers. Therefore the sensitivity and specificity values can not be used as initialization values in the EM-based STAPLE algorithm (Section 3.2). Instead we follow the Bayesian Decision framework and calculate directly $f(T_i=1|D)$ using Eqs. 10, 11, and 12 with the known (p, q) values of observers; the unknown truth prior probability is modeled through one of two ways:

A) We assume there is no prior available about the ground truth map and initialize with uniform distribution (i.e. $f(T_i=1)=f(T_i=0)=0.5$).



(a) (b) (c)

Figure 12 Results for the experiments in which p and q values are known for some observers. **a** Original Image. **b** Result for Experiment 1. **c** Result for Experiment 2.

Table 8 Initial (p, q) values with t=0.9999.

Experiment	γ	Value	Observer 1 (red)	Observer 2 (green)	Observer 3 (blue)	Result
Experiment 1	0.3	Initial p	0.9999	0.9999	0.9999	Fig. 13b
		Initial q	0.9999	0.9999	0.9999	
		Final p	0.893	0.983	0.986	
		Final q	0.946	0.971	0.969	
Experiment 2	0.5	Initial p	0.9999	0.9999	0.9999	Fig. 13c
		Initial q	0.9999	0.9999	0.9999	
		Final p	0.893	0.983	0.986	
		Final q	0.946	0.971	0.969	

B) We assume the observers' segmentation data reflect the prior distribution of the true segmentation and thus initialize the prior probability using the data. (STAPLE adopts this initialization scheme in the absence of truth prior). More specifically, we can either initialize with a single global (homogeneous) prior $\boldsymbol{\gamma}$ as the sample mean of the relative proportion of the label in the multiple observers' segmentations [1]:

$$\gamma = f(T_i = 1) = \frac{1}{RN} \sum_{j=1}^{R} \sum_{i=1}^{N} D_{ij} \tag{13}$$

or with a spatially varying prior map as the sample mean of all observers' labels:

$$f(T_i = 1) = \frac{1}{R} \sum_{j=1}^{R} D_{ij} \tag{14}$$

Sometimes we have the performance measures of some observers but not others. Our approach in this situation is to use the above algorithm to estimate the ground truth, and then the observers with unknown measures are evaluated by comparing their segmentations to the estimated ground truth. The (p, q) values are calculated by Eq. 1 and 2.

4.3 Only Truth Prior Probability $\gamma=f(T_i=1)$ is Known

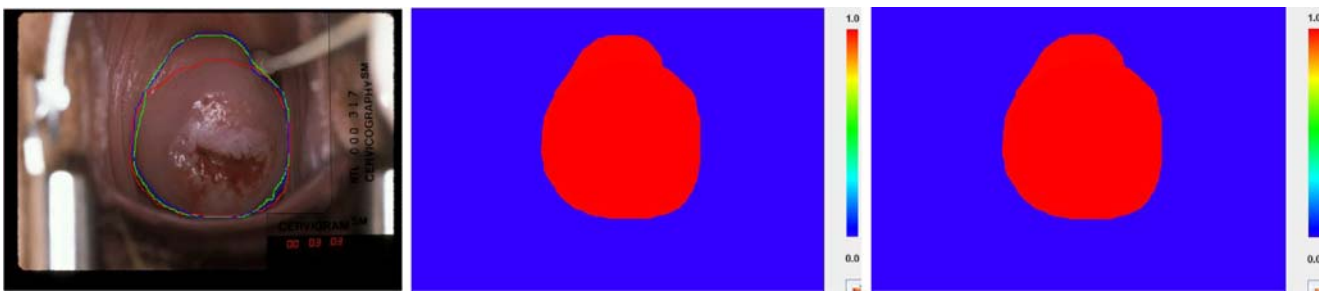
In this case, the known truth prior is directly applied in Eq. 12, while the missing (p, q) values of each observer can be set in two ways:

A) We assume everyone has the same performance level thus the same (p, q) values, i.e., $p_i=q_i=t$ ($0<t<1$). In reality, t can be much smaller than 100%. Whenever this value changes, the estimated ground truth probability map changes accordingly, which reflects the changing confidence in the observers.

B) Similar to Section 4.2B), we can initialize the (p, q) values of each observer based on the multiple observers' segmentation data. In this case, the sample mean map (Eq. 14) is taken as the prior estimate of the ground truth and a threshold of 0.5 is applied to the probability map to obtain a binary map. Then the initial (p, q) values are calculated by using Eq. 1 and 2.

4.4 No Prior Information is Known

In this scenario, initialization of the truth prior probability and the (p, q) values of each observer in the Bayesian framework



(a) (b) (c)

Figure 13 Experiment results for group one in case three: assuming all observers have equal p=q=0.9999. **a** Original Image. **b** Result for Experiment 1. **c** Result for Experiment 2.

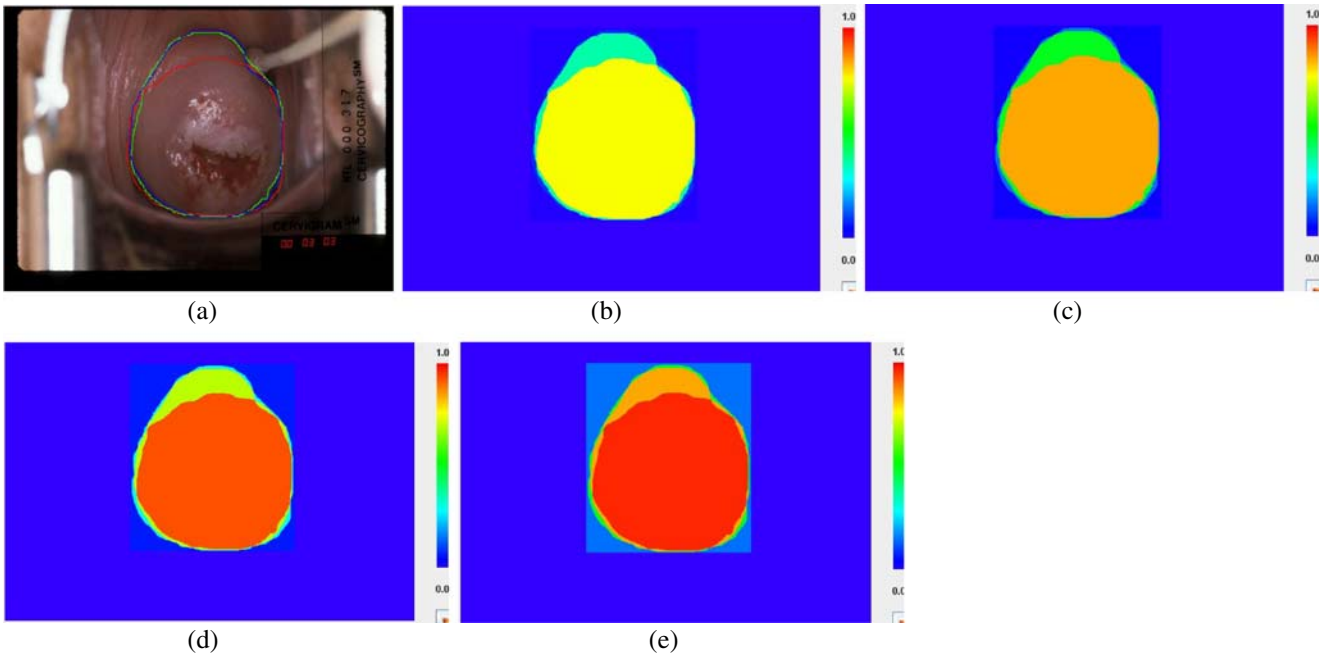


Figure 14 Experiment results for group two in case three: assuming all observers have equal $p=q=0.7$. **a** Original Image. **b** Result for Experiment 1. **c** Result for Experiment 2. **d** Result for Experiment 3. **e** Result for Experiment 4.

is a combination of the initialization methods introduced in Sections 4.2 and 4.3: the initialization of $f(T_i=1)$ (and $f(T_i=0)$) can follow either Sections 4.2A) or 4.2B); the initialization of individual observer's (p, q) values can follow either Section 4.3A) or 4.3B).

5 Software Development

Based on the framework we proposed, we developed a web-based software application. The software is developed in Java and the architecture of the software is shown in Fig. 6.

The system consists of three components: the web browser, the application and the server. The web browser is accessible to users by which they download and evoke the Java application. It is made possible by the Java Web Start technology. The Java application has the following features:

1) Loading and viewing the image and segmentation information. The segmentations of multiple observers are shown on the image in different colors selected

Table 9 Initial (p, q) values with $t=0.7$.

Experiment	γ	Value	Observer 1 (red)	Observer 2 (green)	Observer 3 (blue)	Result
Experiment 1	0.2	Initial p	0.7	0.7	0.7	Fig. 14b
		Initial q	0.7	0.7	0.7	
		Final p	0.9999	0.9999	0.9999	
		Final q	0.899	0.739	0.731	
Experiment 2	0.3	Initial p	0.7	0.7	0.7	Fig. 14c
		Initial q	0.7	0.7	0.7	
		Final p	0.893	0.983	0.986	
		Final q	0.946	0.971	0.969	
Experiment 3	0.5	Initial p	0.7	0.7	0.7	Fig. 14d
		Initial q	0.7	0.7	0.7	
		Final p	0.893	0.983	0.986	
		Final q	0.946	0.971	0.969	
Experiment 4	0.7	Initial p	0.7	0.7	0.7	Fig. 14e
		Initial q	0.7	0.7	0.7	
		Final p	0.877	0.954	0.958	
		Final q	0.9999	0.9999	0.9999	

Table 10 (p, q) values initialized with data.

Experiment	γ	Value	Observer 1 (red)	Observer 2 (green)	Observer 3 (blue)	Result
Experiment 1	0.3	Initial p	0.978	0.990	0.988	Fig. 15b
		Initial q	0.763	0.954	0.961	
		Final p	0.944	0.975	0.972	
		Final q	0.763	0.9999	0.9999	
Experiment 2	0.5	Initial p	0.978	0.990	0.988	Fig. 15c
		Initial q	0.763	0.954	0.961	
		Final p	0.978	0.99	0.989	
		Final q	0.763	0.954	0.962	

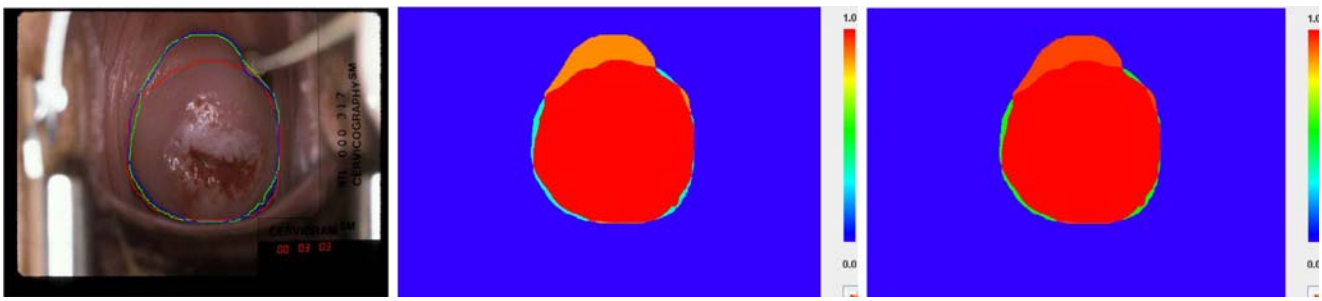
automatically. The detailed information of segmentations is listed in a table format including user names, colors and the initial (p, q) values. In the table, the segmentations can be switched on or off displaying. The color in which a segmented region boundary is drawn can be from a color panel. Figure 7a shows the user interface for loading and viewing the image and segmentations.

2) Communicating with the server and displaying results. A user may select among the different scenarios implemented in our framework: with known (p, q) values for each observer, with known ground-truth prior probability (between 0 and 1), or without any prior knowledge. Furthermore, the application also has an option for computing the combined ground truth map by the Majority Vote rule for comparison purposes. After the user selects an option and sets appropriate prior values, the application submits the image, multiple-observer segmentation and prior information to the server and receives evaluation results from the server (Fig. 7b). The estimated ground truth map is shown on the panel of the application. When clicking a pixel on the map, its position and its probability of being inside the true segmentation are displayed in textboxes. The ground truth map and the original image can also be displayed side-by-side for comparison (Fig. 7c).
3) Exporting the final results including the posterior ground truth map and the (p, q) values (if changed) to files in a selected local directory. The ground truth map is saved in the format of a grayscale image while the final (p, q) values in text format.
4) Quick-start guide. The help documentation for a quick start is developed with JavaHelp 2.0. It allows users to search for keywords in the document.

The software on the server side includes a Java servlet and algorithms. The Java servlet communicates with the application. It receives the image, observer segmentations, and prior information from the application and sends the results back to the application after the algorithms finish computing.

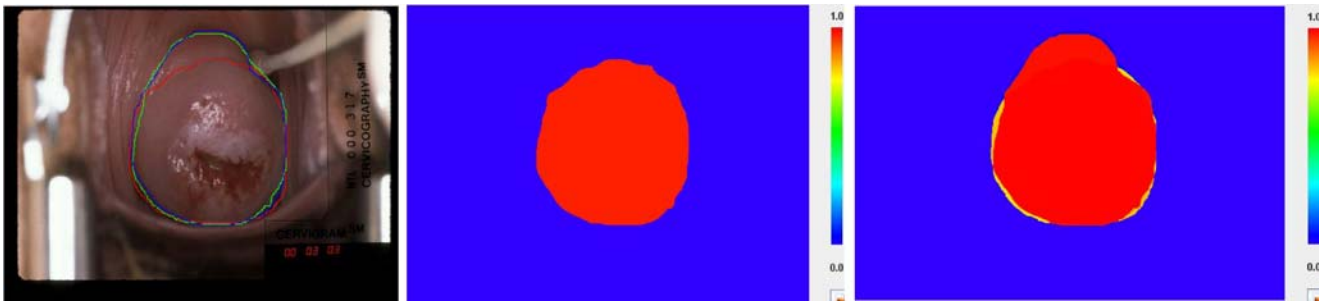
6 Experimental Results Using Manual Segmentations

We carried out several experiments in four scenarios as described in Section 4 by using a subset of images from the NCI/NLM database which contains 939 cervigrams with multi-observer segmentation data. The image is rescaled to half size of the original one which has the size of 2399× 1636 pixels. It is segmented by one to twenty medical experts with varying performance level. For clarity of



(a) (b) (c)

Figure 15 Experiment results for case three with known truth prior and data-initialized (p, q). **a** Original Image. **b** Result for Experiment 1. **c** Result for Experiment 2.



(a) (b) (c)

Figure 16 Estimated ground truth maps with the setups in Table 13. **a** Original Image. **b** Result for Experiment 1. **c** Result for Experiment 2.

presentation, we show our results on one image (Fig. 8a) that was segmented by three observers and compare the results with those by the STAPLE algorithm and the Majority Vote rule. Experimental and comparison results on other images in the database have shown similar trends. On the example image, two observers (in green and blue lines) give similar segmentation while the other (in red line) is different from the two. The sensitivity and specificity values are calculated inside the bounding box of ROI (area of interest) and not in the whole images.

6.1 Scenario One: No Prior Information is Known

6.1.1 Results of Our Method

We initialize the truth prior probability and the observer (p, q) values as outlined in Section 4.4:

A) Assume a single global prior probability γ=0.5 and every observer has equal sensitivity and specificity, i.e. p_i=q_i=t. We choose t=0.9999 and t=0.7 in Experiments 1 and 2 respectively (Table 3; Fig. 8).

In these two experiments, the initial (p, q) values are set differently, and one can see that the estimated ground truth maps indicate changing probability due to changes in observer performance-level priors. It should be noted that although Experiment 2 has different probability map from Experiment 1, it generates the same binary ground truth map as Experiment 1 since we set the probability threshold to distinguish the foreground from background equal to 0.5. Thus the final (p, q) values are the same in these two experiments. If the (p, q) prior values were set to be much lower in Experiment 2, the binary ground truth map and the final (p, q) values would differ from Experiment 1.

B) Use data to initialize the truth prior probability and the (p, q) values of each observer (Table 4; Fig. 9).

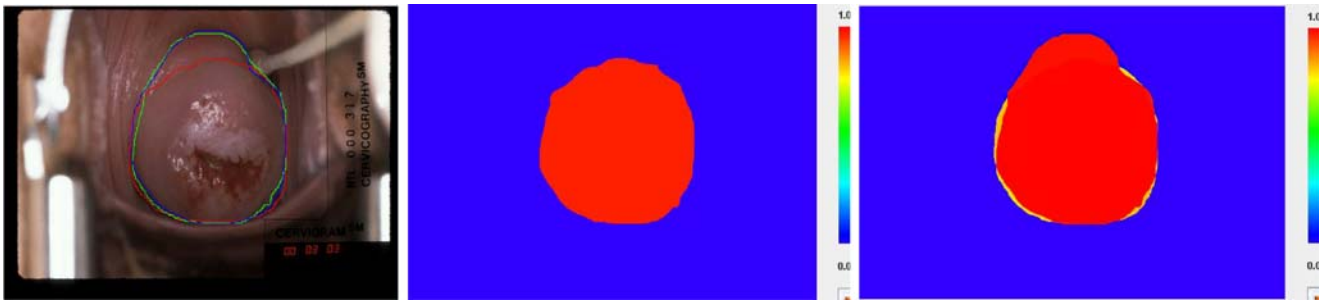
In this case, we initialize the prior probability (Section 4.2B) and (p, q) values (Section 4.3B) based on the observers' segmentation data. The resulting ground truth map (Fig. 9b) is similar to that of the Major Vote Rule shown in Fig. 10d.

6.1.2 Compared to the Results of STAPLE and Majority Vote Rule

In STAPLE, since there is no prior knowledge about either the truth prior probability or (p, q) values of each observer, the prior probability is estimated as the sample mean of the relative proportion of the label in the segmentation (Eq. 3). This means that each observer is treated as equal. Since the truth prior is the dominant prior in the STAPLE algorithm, the results generated by STAPLE are similar to that of the Majority Vote Rule (Fig. 10d). The initial (p, q) values for

Table 11 STAPLE experiments with known truth prior probability and assuming equal (p, q) for each observer: p=q=0.9999.

Experiment	γ	Value	Observer 1 (red)	Observer 2 (green)	Observer 3 (blue)	Result
Experiment 1	0.3	Initial p	0.9999	0.9999	0.9999	Figs. 16b and 17b
		Initial q	0.9999	0.9999	0.9999	
		Final p	0.9999	0.9999	0.9999	
		Final q	0.855	0.703	0.695	
Experiment 2	0.5	Initial p	0.9999	0.9999	0.9999	Figs. 16c and 17c
		Initial q	0.9999	0.9999	0.9999	
		Final p	0.893	0.985	0.998	
		Final q	0.931	0.961	0.959	



(a) (b) (c)

Figure 17 Estimated ground truth maps with the setups in Table 14. **a** Original Image. **b** Result for Experiment 1. **c** Result for Experiment 2. **d** Result for Experiment 3.

each observer have little effect on the results. This can be seen in Fig. 10b and c. The initial (p, q) values of each observer are listed in Table 5.

Using the Majority Vote Rule, the truth prior and the initial (p, q) values are irrelevant and the results are completely determined by majority of the data, which is a shortcoming of the rule (Table 5; Fig. 10).

6.2 Scenario Two: Only Observer (p, q) Values are Known

6.2.1 Results of Our Method

In the first group of experiments, we consider the case in which the (p, q) prior values for all observers are known (Table 6). In Experiment 1, each observer has equal (p, q) values while in Experiment 2, Observer 1 is an expert and Observers 2 and 3 are non-experts. The results are consistent with the (p, q) values set for each observer. In Experiment 2, the result leans toward the segmentation by Observer 1, who is an expert (Fig. 11c; Table 6).

The other situation in this scenario can be that the measures of performance are known for some observers only. In the second group of experiments, we specify (p, q) values for some observers (Table 7). The other observers are evaluated by our framework (Section 4.2) and their measures of performance are shown in Table 7 (Fig. 12).

6.2.2 Compared to the Results of STAPLE and Majority Vote Rule

As discussed in Section 3.2, the limitation of the STAPLE algorithm is that the truth prior probability is dominant and the (p, q) prior values of each observer are ignored (see Table 1, 2, 5 and Fig. 4, 5, 10). Thus the STAPLE algorithm does not apply to this scenario. The Majority Vote Rule generates results that depend on observer data alone without considering prior information so it does not apply to this scenario either.

6.3 Scenario Three: Only Truth Prior Probability $\gamma = f(T_i = 1)$ is Known

6.3.1 Results of Our Framework

We initialize (p, q) values for each observer as outlined in Section 4.3:

A) Assume every observer has equal sensitivity and specificity, i.e. $p_i = q_i = t$. In order to see the effect of the prior probability and (p, q) values for each observer, we carried out two groups of experiments. In one group, we set $t = 0.9999$, and in the other, $t = 0.7$. In each group of experiments, we also changed γ between 0.2 and 0.7.

Table 12 STAPLE experiments with known truth prior probability and assuming equal (p, q) for each observer: $p = q = 0.7$.

Experiment	γ	Value	Observer 1 (red)	Observer 2 (green)	Observer 3 (blue)	Result
Experiment 1	0.3	Initial p	0.7	0.7	0.7	Figs. 18b
		Initial q	0.7	0.7	0.7	and 19b
		Final p	0.9999	0.9999	0.9999	
		Final q	0.855	0.703	0.695	
Experiment 2	0.5	Initial p	0.7	0.7	0.7	Figs. 18c
		Initial q	0.7	0.7	0.7	and 19c
		Final p	0.893	0.985	0.998	
		Final q	0.931	0.961	0.959	

Table 13 The prior probability and (p, q) values of each observer for experiments in group one.

Experiment	γ	Value	Observer 1 (red)	Observer 2 (green)	Observer 3 (blue)	Result
Experiment 1	0.3	Initial p	0.9999	0.9999	0.9999	Figs. 16b and 17b
		Initial q	0.9999	0.9999	0.9999	
		Final p	0.893	0.983	0.986	
		Final q	0.946	0.971	0.969	
Experiment 2	0.5	Initial p	0.9999	0.9999	0.9999	Figs. 16c and 17c
		Initial q	0.9999	0.9999	0.9999	
		Final p	0.893	0.983	0.986	
		Final q	0.946	0.971	0.969	

In the first group of experiments (Table 8), each observer has high sensitivity and specificity thus their effect overwhelms the effect of the prior probability (Fig. 13 and 14; Table 8 and 9).

In the second group of experiments (Table 9), each observer is initialized with lower sensitivity and specificity so we clearly see the effect of the truth prior probability.

Therefore, it is recommended that when there is reliable information about the truth prior but no knowledge about observer performance levels, a small t value be used to initialize the (p, q) values of each observer.

B) Use observers' segmentation data to initialize (p, q) values (Table 10; Fig. 15).

In this group of experiments (Table 10), each observer has initial sensitivity and specificity calculated from the segmentation data. We clearly see the effect on the estimated ground truth probability map given changes in the truth prior probability.

6.3.2 Results of STAPLE

In order to compare our results with those from the STAPLE algorithm, we applied STAPLE with the same configurations as in Table 8 and 9.

In the first group of experiments, as one can see, the prior probability has a significant effect and the results are consistent with the prior probability (Fig. 16; Table 11).

In the second group of experiments, we set (p, q) values for each observer lower. At the same time, we change the prior probability. The results show again that the truth prior dominates over the observer prior (p, q) in STAPLE (Fig. 17). By comparing the first and second groups of experiments, one can see that the (p, q) settings do not affect STAPLE's final results. For instance, the resulting ground truth map Fig. 16b is exactly the same as Fig. 17b, and Fig. 16c the same as Fig. 17c, even though the (p, q) values in these two groups of experiments are very different. This again shows STAPLE's limitation pointed out in Section 3.2 (Table 12; Fig. 17).

6.4 Scenario Four: Both Truth Prior Probability $\gamma=f(T_i=1)$ and Observer (p, q) Values are Known

6.4.1 Results of Our Method

When we have reliable estimates of both the truth prior probability and observer (p, q) values, our method coherently balances their effects and integrates them in a complementary manner. We carried out two groups of experiments in this case. One is with higher (p, q) values

Table 14 The prior probability and (p, q) values of each observer for the experiments of group two.

Experiment	γ	Value	Observer 1 (red)	Observer 2 (green)	Observer 3 (blue)	Result
Experiment 1	0.1	Initial p	0.9999	0.9999	0.9999	Figs. 18b and 19b
		Initial q	0.9999	0.7	0.7	
		Final p	0.9999	0.9999	0.9999	
		Final q	0.899	0.739	0.731	
Experiment 2	0.3	Initial p	0.9999	0.9999	0.9999	Figs. 18c and 19c
		Initial q	0.9999	0.7	0.7	
		Final p	0.9999	0.981	0.984	
		Final q	0.958	0.751	0.75	
Experiment 3	0.5	Initial p	0.9999	0.9999	0.9999	Figs. 18d and 19d
		Initial q	0.9999	0.7	0.7	
		Final p	0.9999	0.981	0.984	
		Final q	0.958	0.751	0.75	

for each observer and the other is with lower (p, q) values for each observer. At the same time, we changed the value of the truth prior probability. As one can see, when the (p, q) values are very high (close to 1.0), the effect of the observer data dominates over the truth prior probability, while when the (p, q) values are lower indicating low confidence in observer data, the truth prior clearly shows its effect (Table 13 and 14; Fig. 16 and 17).

The STAPLE algorithm cannot handle this Scenario since the truth prior probability dominates even with very high *(p, q)* values for each observer.

7 Experimental Results Using an Automatic Segmentation Method

In this experiment, we use our framework to evaluate our automatic segmentation method [20]. First, we use the automatic segmentation method to differentiate the acetowhite (AW) issue and non-AW tissue. Then we use the results from multiple observers' manual segmentations to evaluate the result from the automatic segmentation method.

7.1 Automatic Classification Using Cluster Features for Lesion Detection in Digital Cervigrams

We use a database-guided segmentation paradigm in which we apply machine learning techniques, such as support vector machines (SVM) to learn, from a database with ground truth annotations provided by experts, critical visual signs that correlate with important tissue types and to use the learned classifier for tissue segmentation in unseen images. The support vector machines (SVM) classifier has been successfully applied to detecting Microcalcifications in Mammograms and various other medical classification problems. We use SVM to perform color-based tissue classification in order to segment different tissue regions, especially to segment the biomarker AW region from the rest of the cervix. The segmentation performance is optimized with respect to the feature color space and granularity. We evaluate color spaces including RGB, HSV, and $L^*a^*b^*$. On different granularity of the features, we train AW and other tissue classifiers, first using individual pixel sample colors and then using cluster features returned by the Mean Shift based clustering algorithm. Cluster features greatly reduce the dimensionality of training so that SVM is scalable to larger training sets, while producing results with comparable accuracy. Given a novel test image, the Mean Shift clustering algorithm partitions the image into clusters of similar color and/or texture, and the trained SVM classifier (on cluster features of training data) is applied to classifying clusters in the test image. This ground-truth database guided segmentation method is flexible in terms of the number of tissue classes. Thus we can perform either two-label, or multi-label classification.

7.2 Results

We demonstrate our results in one scenario where no prior information is known. We use the segmentation data for initializing the unknown priors: the probability prior and the (p, q) values of multiple observers. Table 15 shows the prior probability and (p, q) values of multiple observers while Fig. 18 shows the original image, the result from our automatic segmentation method and the ground truth map. In Experiment 1 and 2, our automatic method has lower sensitivity than specificity partly because the automatic method excluded the os part of the cervix (Table 15; Fig. 18).

Table 15 The prior probability and (p, q) values of each observer for the experiments.

Experiment	γ	Value	Observer 1 (red)	Observer 2 (green)	Observer 3 (blue)	Automatic	Result
Experiment 1	0.53	Initial p	0.9999	0.9999	0.9999	N/A	Fig. 20 *1*c
		Initial q	0.9999	0.9999	0.9999	N/A	
		Final p	0.902	0.979	0.964	0.74	
		Final q	0.99	0.965	0.861	0.865	
Experiment 2	0.52	Initial p	0.9999	0.9999	0.9999	N/A	Fig. 20 *2*c
		Initial q	0.9999	0.9999	0.9999	N/A	
		Final p	0.93	0.984	0.971	0.873	
		Final q	0.995	0.954	0.92	0.929	
Experiment 3	0.44	Initial p	0.9999	0.9999	0.9999	N/A	Fig. 20 *3*c
		Initial q	0.9999	0.9999	0.9999	N/A	
		Final p	0.995	0.843	0.754	0.805	
		Final q	0.669	0.97	0.954	0.801	

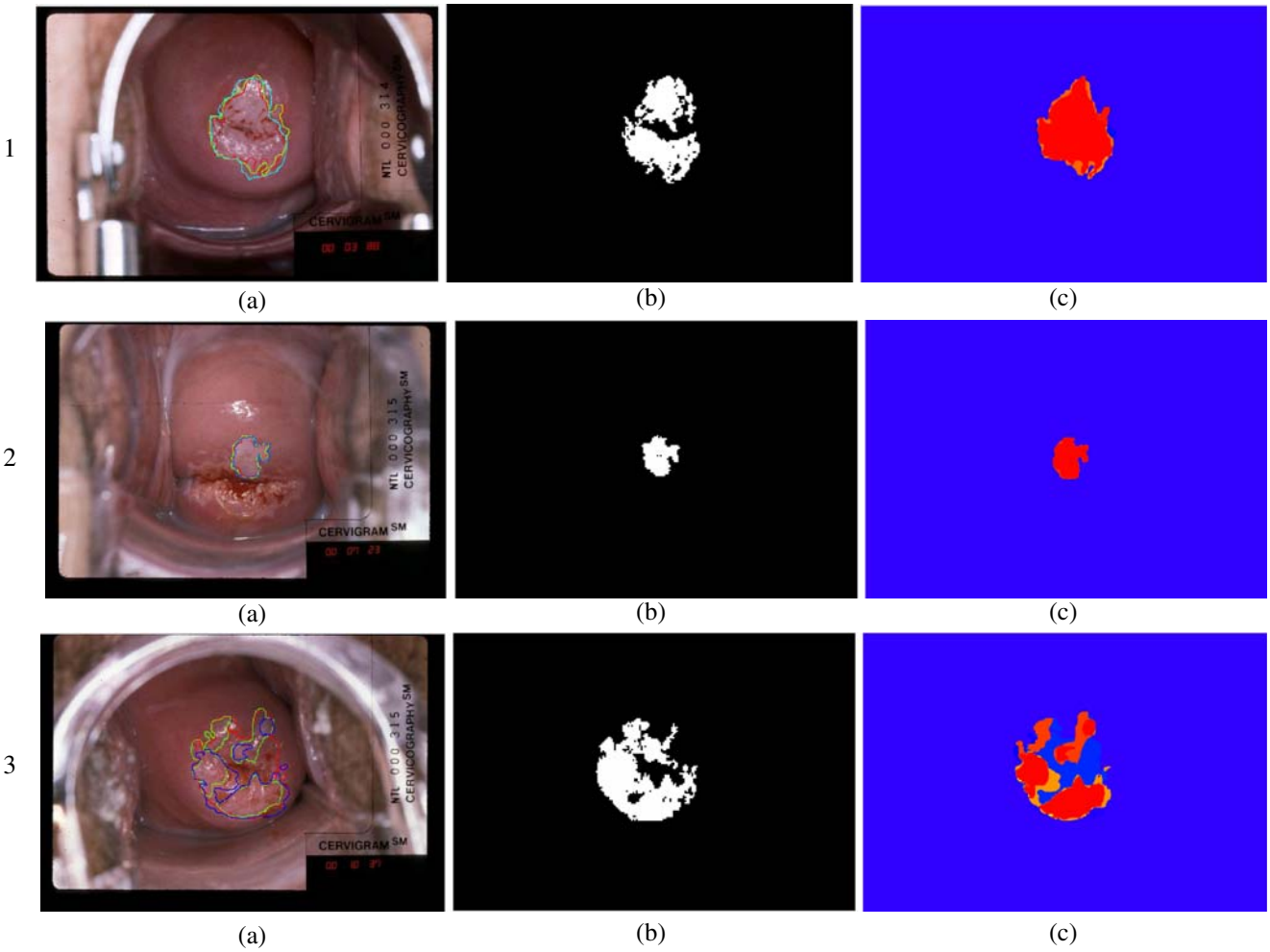


Figure 18 Estimated ground truth maps with the setups in Table 15. *1***a** Original Image. **b** Result for Our Automatic Segmentation Method. **c** Result for Experiment 1. *2***a** Original Image. **b** Result for Our Automatic Segmentation Method. **c** Result for Experiment 2. *3***a** Original Image. **b** Result for Our Automatic Segmentation Method. **c** Result for Experiment 3.

8 Conclusion and Future Work

In this paper, we have proposed a new method for multiple observer segmentation evaluation based on analysis of the STAPLE algorithm. The analysis includes different scenarios that have different kinds of prior knowledge available. We first identified a limitation of the STAPLE algorithm which indicates that observer performance prior is effectively ignored in the framework. We formulate instead a Bayesian Decision framework that balances the roles of the ground truth segmentation prior and observer performance-level prior according to their availability and confidence in their estimation. We demonstrate multi-observer segmentation evaluation results of our framework in four scenarios with differing prior knowledge and application purposes, and the results compare favorably to those by the STAPLE algorithm and the Majority Vote Rule. The results also show the flexibility of our method in effectively integrating different priors for multi-observer segmentation evaluation. Although we only illustrate the results by using the cervigrams, our method can work for multi-observer segmentation applications using any images. Currently, our online software only allows users to submit the segmentation information to the server in the format of contours in order to save the transfer time. We will extend the software to include binary images and other formats in the future. Another missing part of our framework is to integrate the constraints such as structure or shape constraints since integration of more prior information will help to generate more accurate evaluation results.

Future work also includes the following directions: (a) the extension of our framework to multiple labels, (b) the extension of our framework to 3D, which is pretty straightforward. The voxels are used instead of pixels. The ground truth map becomes a 3D probability map. All equations in our framework remain the same as those in

2D. (c) similar to that in STAPLE, our framework can take the spatial prior into consideration, (d) the current method only works on a single image with multiple observers' segmentations. It can be extended to evaluate each observer's performance based on their segmentations on multiple images, (e) we plan to apply this method to evaluating the performance of automatic segmentation algorithms and to improving the consensus in training, and (f) the method can be integrated in model-based segmentation frameworks to provide feedback on how to refine model parameters.

References

1. Warfield, S. K., Zou, K. H., & Wells, W. M. (2004). Simultaneous Truth and Performance Level Estimation (STAPLE): An algorithm for the validation of image segmentation. *IEEE Transactions on Medical Imaging*, July.
2. Lotenberg, S., Greenspan, H., Gordon, S., Long, L. R., Jeronimo, J., & Antani, S. K. (2007). Automatic evaluation of uterine cervix segmentations. *Proceedings of SPIE Medical Imaging*, *6515*, 65151J–1-12.
3. Zhu, Y., Long, L. R., Antani, S. K., Xue, Z., & Thoma, G. R. (2007). Web-based STAPLE for quality estimation of multiple image segmentations. Poster at 20th NIH Research Festival (IMAG-12), National Institutes of Health, September.
4. Zhang, Y. J. (1996). A survey on evaluation methods for image segmentation. *Pattern Recognition*, *29*(8), 1335–1346.
5. Yasnoff, W. A., Mui, J. K., & Bacus, J. W. (1977). Error measures in scene segmentation. *Pattern Recognition*, *9*(4), 217–231.
6. Qian Huang Dom, B. (1995). Quantitative methods of evaluating image segmentation. *Proceedings IEEE International Conference on Image Processing*, *3*, 53–56.
7. Martin, D. (2002). An empirical approach to grouping and segmentation. PhD dissertation, University of California, Berkeley.
8. Cardoso, J. S., & Corte-Real, L. (2005). Toward a generic evaluation of image segmentation. *IEEE Transactions on Image Processing*, *14*(11), 1773–1782.
9. Monteiro, F. C., Fernando, C., Campilho, A. C., & Aurélio, C. Performance Evaluation of Image Segmentation. ICIAR06 (I: 248–259).
10. Kittler, J., Hatef, M., Duin, R. P. W., & Matas, J. (1998). On combining classifiers. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *20*, 226–239 Mar.
11. Windridge, D., & Kittler, J. (2003). A morphologically optimal strategy for classifier combination: Multiple expert fusion as a tomographic process. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *25*, 343–353 Mar.
12. Jacobs, R. A., Jordan, M. I., Nowlan, S. J., & Hinton, G. E. (1991). Adaptive mixtures of local experts. *Neural Computation*, *3*, 79–87.
13. Jordan, M. I., & Jacobs, R. A. Hierarchical Mixtures of Experts and the EM Algorithm. Tech. Rep. AIM-1440, 1993.
14. Restif, C. (2007). Revisiting the evaluation of segmentation results: Introducing confidence maps. *Medical Image Computing and Computer-Assisted Intervention*, *2*, 588–595.
15. Martina, A., Laanaya, H., & Arnold-Bos, A. (2006). Evaluation for uncertain image classification and segmentation. *Pattern Recognition*, *39*(11), 1987–1995 November.
16. Berger, J. (1985). *Statistical decision theory and bayesian analysis*. New York: Springer-Verlag.
17. Prasad, M., Sowmya, A., & Koch, I. (2004). Feature subset selection using ICA for classifying emphysema in HRCT images. *17th International Conference on Pattern Recognition (ICPR)*, *4*, 515–518.
18. Prasad, M., Sowmya, A., & Wilson, P. Multi-level classification of emphysema in HRCT lung images. Pattern Analysis & Applications
19. Herrero, R., Schiffman, M. H., Bratti, C., et al. (1997). Design and methods of a population-based natural history study of cervical neoplasia in a rural province of Costa-Rica: The Guanacaste Project. *Revista Panamericana de Salud Pública*, *1*(5), 362–375.
20. Huang, X., Wang, W., Xue, Z., Antani, S., Long, L. R., & Jeronimo, J. (2008). *Tissue classification using cluster features for lesion detection in digital cervigrams*. San Diego: SPIE Medical Imaging.

Yaoyao Zhu is currently a Ph.D. candidate in the Computer Science and Engineering Department at Lehigh University. She received a B.S. degree in electronics from Beijing University, China, in 1995 and an M.S. degree in computer engineering from the University of Cincinnati in 2001. Her research interests include machine learning, pattern recognition and medical image processing.

Xiaolei Huang received her doctorate and masters degrees in computer science from Rutgers, the State University of New Jersey in 2006 and 2001 respectively, and her bachelors degree in computer science from Tsinghua University, China, in 1999. Since 2006, she has been an Assistant Professor in the Computer Science and Engineering Department at Lehigh University. Her research interests include Computer Vision, Biomedical Image Analysis, and Computer Graphics. In these areas, she has authored or co-authored over 30 publications including journal articles, book chapters, and refereed

conference proceedings papers. A member of the Institute of Electrical and Electronics Engineers and the Biomedical Engineering Society, she has served on the program committees of several biomedical imaging and computer vision conferences and reviews papers for journals including IEEE Transactions on Pattern Analysis and Machine Intelligence, Graphical Models, Medical Image Analysis, and IEEE Transactions on Biomedical Engineering. She is the holder of 1 U.S. patent and has 5 U.S. patents pending.

Wei Wang received the BS degree in electronics and information engineering from Beihang University (Beijing University of Aero and Astro) in 2003 and the MS degree in electrical engineering from Lehigh University in 2007. He is currently a Ph.D. candidate work at IDEA lab in Lehigh University. His research interests are clustering, image segmentation and medical application.

Daniel Lopresti received his bachelors degree from Dartmouth College, Hanover, NH in 1982 and his Ph.D. degree in computer science from Princeton University, Princeton, NJ in 1987. He spent several years with the Computer Science Department, Brown University, Providence, RI, and then went on to help found the Matsushita Information Technology Laboratory in Princeton. He later spent time at Bell Labs. Since 2003, he has been with the Computer Science and Engineering Department, Lehigh University, Bethlehem, PA, where he leads research examining fundamental algorithmic and systems-related questions in pattern recognition, document analysis, bioinformatics, and computer security. He has authored or co-authored over 100 publications in journals and refereed conference proceedings and is the holder of 21 U.S. patents.

Rodney Long is an electronics engineer for the Communications Engineering Branch at the National Library of Medicine, where he has worked since 1990. Prior to his current job, he worked for 14 years in industry as a software developer and as a systems engineer. His research interests are in telecommunications, image processing, and scientific/biomedical databases. He has an M.A. in applied mathematics from the University of Maryland. He is a member of the Mathematical Association of America and the IEEE.

Dr. Sameer Antani is a Staff Scientist with the Lister Hill National Center for Biomedical Communications an intramural R&D division of the National Library of Medicine (NLM) at the U.S. National Institutes of Health (NIH). His research interests are in image and text data management for large biomedical and multimedia archives. His research includes content-based indexing, and retrieval of biomedical images (CBIR), combining image and text retrieval, topics in advanced multimodal medical document retrieval, and next-generation interactive (multimedia rich) documents. He earned his B.E. (Computer) degree from the University of Pune, India, in 1994, and his M.E. and Ph.D. degrees in Computer Science and Engineering from the Pennsylvania State University, USA, in 1998 and 2001, respectively. Dr. Antani is a member of the IEEE, the IEEE Computer Society, and SPIE. He serves

on the steering committee for IEEE Symposium for Computer Based Medical Systems (CBMS).

Zhiyun Xue joined the Lister Hill National Center for Biomedical Communications at the National Library of Medicine (NLM) in 2006. Her research interests are in the areas of medical image analysis, computer vision, and pattern recognition. She received her Ph.D. degree in Electrical Engineering from Lehigh University in 2006, and her master's and bachelor's degrees in Electrical Engineering from Tsinghua University, China, in 1998 and 1996, respectively.

George R. Thoma received the B.S. from Swarthmore College, and the M.S. and Ph.D. from the University of Pennsylvania, all in electrical engineering. As the senior electronics engineer and Chief of the Communications Engineering Branch of the Lister Hill National Center for Biomedical Communications, a research and development division of the National Library of Medicine, he directs R&D programs in image processing, document image storage on digital optical disks, automated document image delivery, digital xray archiving, and high speed image transmission. He has also conducted research in analog videodiscs, satellite communications and video teleconferencing. Dr. Thoma is a Fellow of the SPIE, the International Society for Optical Engineering.